A NOVEL SCHEME FOR SPEAKER RECOGNITION USING A PHONETICALLY-AWARE DEEP NEURAL NETWORK

Yun Lei *Nicolas Scheffer* *Luciana Ferrer* *Mitchell McLaren*

Speech Technology and Research Laboratory, SRI International, California, USA

{yunlei,scheffer,lferrer,mitch}@speech.sri.com

ABSTRACT

We propose a novel framework for speaker recognition in which extraction of sufficient statistics for the state-of-the-art i-vector model is driven by a deep neural network (DNN) trained for automatic speech recognition (ASR). Specifically, the DNN replaces the standard Gaussian mixture model (GMM) to produce frame alignments. The use of an ASR-DNN system in the speaker recognition pipeline is attractive as it integrates the information from speech content directly into the statistics, allowing the standard backends to remain unchanged. Improvement from the proposed framework compared to a state-of-the-art system are of 30% relative at the equal error rate when evaluated on the telephone conditions from the 2012 NIST speaker recognition evaluation (SRE). The proposed framework is a successful way to efficiently leverage transcribed data for speaker recognition, thus opening up a wide spectrum of research directions.

Index Terms— deep neural network, speaker recognition

1. INTRODUCTION

Recently, the speaker verification community has seen a significant increase in accuracy from the successful application of the i-vector extraction paradigm [1]. This framework can be decomposed into three sequential stages: the collection of sufficient statistics, the extraction of i-vectors and a probabilistic linear discriminant analysis (PLDA) backend. The collection of sufficient statistics is a process where a sequence of feature vectors (e.g., mel-frequency cepstral coefficients (MFCC)) are represented by the Baum-Welch statistics obtained with respect to a GMM, refered to as universal background model (UBM). These statistics are highly dimensional, and are converted into a single low-dimensional feature vector — an i-vector — that represents important information about the speaker and all other types of variability in a given speech segment. Once i-vectors are extracted, a PLDA model is then used to produce verification scores by comparing i-vectors extracted from different utterances [2].

In the field of speech recognition, deep neural networks (DNN) have recently been successfully used for acoustic modeling, achieving large improvements compared to standard GMM models [3, 4]. The DNN is a standard feed-forward neural network that is both much larger (a few thousand nodes per hidden layer) and much deeper (roughly 5-7 hidden layers) than traditional neural networks. Standard discriminative back-propagation algorithm and stochastic gradient descent approach are typically used for the DNN training.

While the application of DNNs in other speech-related fields is straightforward (each output node of the DNN represents one of the classes of interest), a direct transition to speaker recognition is much more challenging, as speakers are often unknown during system training and each speaker has very little training data.

Our work aims to use a DNN trained for speech recognition to guide speaker modeling, specifically, by using the output posteriors as frame alignments for speaker modelling and i-vector extraction, substituting for the role of the UBM in the standard framework. Our use of a phonetically aware model is motivated by the fact that the speech content's effect on the speech signal have been mostly ignored in work on text-independent speaker verification. Prior studies on phone-, syllable- or word-dependent GMM systems [5, 6, 7, 8] or constrained systems [9] have shown promise but are not widely adopted due to their complexity and the marginal improvements in accuracy they provide, even after combination with a baseline system. A content-aware system, efficiently leveraging transcribed data, opens up a wide spectrum of possibilities for research and improvements in speaker recognition.

In this work, the DNN replaces the GMM to compute the posterior of the frames with respect to each of the classes in the model. While in the case of the GMM, the classes are the individual Gaussians from a mixture model, in the case of the DNN, the classes are senones (tied triphone states) obtained using a standard decision tree for automatic speech recognition. Once the posteriors are computed, the zeroth and first order statistics are computed in the standard way before they are fed into the state-of-the-art paradigm i-vector / PLDA. An attractive benefit of our proposed approach is that the features used for frame alignments and the sufficient statistics can be different, as the two processes are now effectively decoupled. As a result, the system can use optimal features to maximize phone recognition accuracy for the frame alignments while using optimal features for speaker recognition to compute the sufficient statistics that are used to obtain the i-vectors and the final speaker verification scores.

We first present the i-vector model, then briefly highlight the roles of the UBM in speaker recognition. We then describe our DNN approach before presenting results and conclusions.

2. THE I-VECTOR MODEL

In the i-vector model [1], the t-th speech frame $\mathbf{x}_t^{(i)}$ from the i-th speech segment is assumed to be generated by the following distri-

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its contracting agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. "A" (Approved for Public Release, Distribution Unlimited)

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Novel Scheme for Speaker Recognition Using a Phonetically-Aware Deep Neural Network	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,Speech Technology and Research Laboratory,333 Ravenswood Avenue,Menlo Park,CA,94025**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We propose a novel framework for speaker recognition in which extraction of sufficient statistics for the state-of-the-art i-vector model is driven by a deep neural network (DNN) trained for automatic speech recognition (ASR). Specifically, the DNN replaces the standard Gaussian mixture model (GMM) to produce frame alignments. The use of an ASR-DNN system in the speaker recognition pipeline is attractive as it integrates the information from speech content directly into the statistics, allowing the standard backends to remain unchanged. Improvement from the proposed framework compared to a state-of-the-art system are of 30% relative at the equal error rate when evaluated on the telephone conditions from the 2012 NIST speaker recognition evaluation (SRE). The proposed framework is a successful way to efficiently leverage transcribed data for speaker recognition, thus opening up a wide spectrum of research directions.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**5**	

bution:

$$\mathbf{x}_t^{(i)} \sim \sum_k \gamma_{kt}^{(i)} \mathcal{N}(\boldsymbol{\mu}_k + \mathbf{T}_k \boldsymbol{\omega}^{(i)}, \boldsymbol{\Sigma}_k) \quad (1)$$

where the $\mathbf{T}_k$ matrices describe a low-rank subspace (called total variability subspace) by which the means of the Gaussians are adapted to a particular speech segment, $\boldsymbol{\omega}^{(i)}$ is a segment-specific standard normal-distributed latent vector, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance of the k-th Gaussian, and $\gamma_{kt}^{(i)}$, as another inputs of the i-vector model, are the alignments of $\mathbf{x}_t^{(i)}$. In general, we represent the alignments by the posterior of the k-th Gaussian, given by:

$$\gamma_{kt}^{(i)} = p(k|x_t^{(i)}) \quad (2)$$

The i-vector used to represent the speech signal is the maximum a posterior (MAP) point estimate of the latent vector $\boldsymbol{\omega}^{(i)}$. It is noted that the alignments can be replaced by the prior (e.g., weights of the UBM) in equation (1). In this case, the variational Bayes inference has to be used in the training and the statistics have to re-collected in every iteration [10].

Equation (1) models a process by which the frame for time t is generated by first choosing a class k according to the distribution given by Equation (2) and then generating the features according to the Gaussian distribution for that class, $\mathcal{N}(\boldsymbol{\mu}_k + \mathbf{T}_k \boldsymbol{\omega}^{(i)}, \boldsymbol{\Sigma}_k)$. Note that the classes can be defined in any way subject to the theoretical restriction that the classes have a Gaussian distribution.

Given a speech segment, the following sufficient statistics can be computed using the posterior probabilities of the classes:

$$\begin{aligned} \mathbf{N}_k^{(i)} &= \sum_t \gamma_{kt}^{(i)} \\ \mathbf{F}_k^{(i)} &= \sum_t \gamma_{kt}^{(i)} x_t^{(i)} \\ \mathbf{S}_k^{(i)} &= \sum_t \gamma_{kt}^{(i)} x_t^{(i)} x_t^{(i)^T} \end{aligned} \quad (3)$$

These sufficient statistics are all that is needed to train the subspace $\mathbf{T}$ and extract the i-vector $\boldsymbol{\omega}^{(i)}$. Note that means and covariances in Equation (1) can be updated during the subspace training process, though this is not necessary to achieve good performance. The reader can refer to [11] for more details.

3. ROLES OF THE UBM

The idea of a universal background model (UBM), represented by a Gaussian mixture model (GMM) trained on many different speakers, has been used in speaker recognition for many years as one of the fundamental components across different frameworks. In the GMM-UBM [12] framework, the UBM is used to derive the speaker-specific model by adapting its means to the speaker's data using a MAP approach. The likelihood ratio of the test data given the UBM and speaker-specific GMM is used for speaker recognition. In the GMM-support vector machine (SVM) [13] framework, the speaker-specific GMM is obtained just as in the GMM-UBM framework, but the means of the GMM are concatenated to generate a vector called a *supervector*, which is then input to the SVM. In the joint factor analysis (JFA) [11] framework, the UBM is used to compute the frame alignments of an utterance to further generate its sufficient statistics.

In the i-vector framework, the standard approach uses the Gaussians in the UBM as the classes k in Equation (1). This approach ensures that the Gaussian approximation for each class is satisfied (by definition) and provides a simple way to compute the posteriors needed to compute the i-vectors: the likelihood of each Gaussian is computed and Bayes rule is used to convert them into posteriors. As explained in Section 2, only the posteriors of each frame for all classes are needed to compute the i-vectors. This suggests that any kind of model can be used to replace the Gaussian mixture model if it defines K classes and can provide a posterior probability for a class given a frame. In this work, we propose to replace the UBM-GMM by a deep neural network (DNN) trained for ASR,

4. DNNS FOR ASR

In state-of-the-art ASR systems the pronunciations of all words are represented by a sequence of senones $\mathcal{Q}$ (e.g., the tied-triphone states). Each senone is used to model the tied states of a set of triphones that are close in acoustic space. In general, the senone set $\mathcal{Q}$ is automatically defined by a decision tree using the maximum likelihood (ML) approach [14]. The decision tree is grown by asking a set of locally optimal questions that give the largest likelihood increase, assuming that the data on each side of the split can be modeled by a single Gaussian. The leaves of the decision tree are then taken as the final set of senones.

Once the set of senones is defined, a Viterbi decoder is used to align the training data into the corresponding senones. These alignments are used to estimate the observation probability distribution $p(x|q)$, where x is an observation vector in the training data and q is the senone. The estimation of the observation probability distribution and the realignment can be optimized alternatively and iteratively. Traditionally, a GMM was used to model this distribution. In recent systems, a DNN is used to estimate the senone posteriors of the acoustic features. The observation probability can be obtained from the posteriors and priors of the senones using Bayes rule, as follows:

$$p(x|q) = p(q|x)p(x)/p(q), \quad (4)$$

where $p(x|q)$ is the observation probability needed for decoding, $p(q)$ is the senone prior and $p(q|x)$ is the senone posterior obtained from the DNN. Figure 1 presents a flow diagram for training a DNN for ASR. A pre-trained hidden markov model (HMM) ASR system with GMM states is needed to generate alignments for the subsequent DNN training. The final acoustic model is composed of the original HMM from the previous HMM-GMM system and the new DNN.

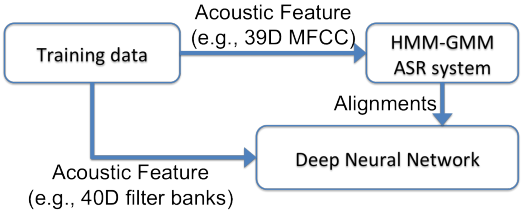

Fig. 1. The flow diagram for training a DNN for ASR.

5. A DNN/I-VECTOR FRAMEWORK

We propose to use the classes k in Equation (1) as the senones defined by the ASR decision tree. (instead of the Gaussian indices in

a GMM), By doing this, we make the assumption that each of these senones can be accurately modeled by a single Gaussian. While a strong assumption, results of this work show that it is a reasonable one for the speaker recognition task.

The motivation behind defining the classes to model the phonetic content is as follows. The i-vector for a certain speech signal models the shifts in means for each class k that are needed to maximize the likelihood of the model (Equation 1) for this signal. The UBM-defined classes and posteriors have no inherent meaning. Each Gaussian simply covers a part of the feature space which might include instances of different phones (or triphones) rather than a single one, or even only some specific pronunciations of a certain phoneme. If a speaker pronounces a certain phoneme, say /aa/, very differently from the general population, the frames corresponding to this phoneme may possibly be aligned with Gaussians trained with other phonemes, say /ao/. As a consequence the shift needed to adapt the Gaussians corresponding to /aa/ will not be affected by the frames in which the speaker was pronouncing /aa/. Only the means for the Gaussians corresponding to /ao/ will be affected by these frames. The final i-vector will then not contain information about the fact that this speaker pronounces /aa/ very differently from others.

On the other hand, when the classes correspond to phonetic senones and the posteriors are accurately computed to predict these senones, the correct frames are used to estimate the shift in the means for each senone. In the example above, the frames corresponding to /aa/ would be assigned to the correct senone and a large shift in the means will result for those senones. The i-vector will then reflect the fact that this speaker pronounces /aa/ very differently from the general population. Simply put, when the classes are defined by the phonetic senones we are able to compare "apples to apples": each frame is compared with the training frames for the same phonemic content.

One could argue that a similar benefit could be achieved by training a UBM in a supervised fashion on transcribed speech data; that is, a UBM where the Gaussians are given by:

$$\begin{aligned}
\gamma_{kt}^{(i)} &\approx p(k|x_t^{(i)}) \\
\pi_k &= \sum_{i,t} \gamma_{kt}^{(i)}, \\
\boldsymbol{\mu}_k &= \frac{\sum_{i,t} \gamma_{kt}^{(i)} x_t^{(i)}}{\sum_{i,t} \gamma_{kt}^{(i)}}, \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{i,t} \gamma_{kt}^{(i)} x_t^{(i)} x_t^{(i)^T}}{\sum_{i,t} \gamma_{kt}^{(i)}} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T. \qquad (5)
\end{aligned}$$

where the ASR system is used to compute the posteriors for each class k for each frame and π_k is the prior of the class k.

This new supervised UBM could then replace the standard UBM and be used to obtain the posteriors needed for i-vector computation by simply calculating the likelihood of each Gaussian and using Bayes rule to convert them to posteriors. Our experimental results below show that this approach does not lead to significant improvements with respect to using the standard UBM trained without knowledge of the phonetic content. We belive that this is due to the relatively poor accuracy of a GMM-based system for phonetic recognition. If a frame corresponding to a certain senone is assigned to another senone, we are no better off than when using the standard unsupervised UBM.

Note that the feature vectors $x_t^{(i)}$ correspond to the standard MFCC features used for speaker recognition. That is, the UBM, supervised or unsupervised, models features that have been optimized for speaker recognition performance rather than ASR. Since context in these features is provided through the deltas and double deltas, they are not powerful enough to be used for accurately predicting phonetic content when modeled by a simple GMM with only a few thousands of Gaussians. For this reason, the UBM cannot be reliably used to compute the posterior for a certain phonetic class.

To solve this problem we propose to directly use the posteriors from the DNN in the ASR system as the γs in eq 3. In acoustic modeling, DNNs have been shown to outperform GMM-based models by a significant margin, due to the fact that they use longer context windows and are discriminatively trained. As a result, a DNN model gives a much better estimate of the senone posterior than the supervised UBM. Note that an important characteristic of our approach is that one does not have to compromise by designing a feature that works well for both ASR and SID. Indeed, the DNN system can use completely different features from the features used for speaker recognition, as long as it improves the estimate of the posterior probability. This is analogous to using a single, central alignment in a multi-feature SID system as opposed feature-dependent alignments. Figure 2 presents a flow diagram of the proposed DNN/i-vector hybrid framework.

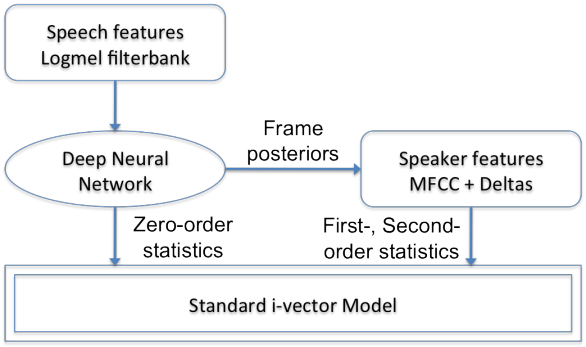

Fig. 2. The flow diagram of the DNN/i-vector hybird farmework.

Much work in the past has pursued the similar goal of comparing matched phonetic content when doing speaker recognition. Several previous studies have investigated constraining or selecting cepstral frames based on word or phone information. For example, the approaches in [5] and [6] condition a cepstral Gaussian mixture model (GMM) on the identities of frequent words or syllables, respectively. The methods described in [7] and [8] assign frames to broad phone classes in order to score them with class-dependent GMMs. Finally, in [9], a constrained cepstral modeling approach was proposed where the constraints were designed to correspond to highly consistent or distinctive phonetic and prosodic features. Most of these approaches are meant to be used in combination with a baseline system since they usually perform worse than the baseline system on their own. Furthermore, they are usually required to make hard decisions about the phones or words found in the signal. These approaches are not widely used as the improvement in accuracy is usually marginal for the increase in complexity.

We believe that the strength of our approach lies in the integration of the phonetic content information into a state-of-the-art baseline system. Rather than designing a system for combination with a baseline, we modify the baseline approach itself to consider this additional source of information. This approach also exposes the system to new data and information: that is, the data and transcripts

used to train the DNN. Furthermore, no hard decisions are made about the phonetic content assignments, thus increasing the robustness of the approach.

6. EXPERIMENTS

The proposed approach is evaluated on the two extended NIST SRE'12 conditions: telephone speech (C2) and telephone speech collected under noisy condition (C5). Since the DNN used in the experiments is trained on a clean English telephone data set, the microphone conditions (e.g., C1 and C3) and noisy telephone condition (C4) are not evaluated in this study. Even though C5 is a noisy condition, the level of noise in those waveforms is known to be significantly lower than for C4. Hence, as we will see, the DNN trained on clean speech is still able to provide good posterior estimates for this condition. Experiments are constrained to female trials to reduce the experimental burden incurred by the state-of-the-art system used in NIST SRE'12.

Both the HMM-GMM and HMM-DNN ASR models are trained on roughly 1300 hours of clean English telephone speech from Fisher, Callhome, and Switchboard data sets. The cross-word triphone HMM-GMM ASR with 3450 senones and 200k Gaussians is trained with maximum likelihood (ML). The features used in the HMM-GMM model are 39-dimensional MFCC features, including 13 static features (including C0) and first and second order derivatives. The features were pre-processed with speaker-based cepstral mean and covariance normalization (MVN). A seven-layer DNN with 600 input nodes, 1200 nodes in each hidden layer and 3450 output nodes was trained with cross entropy using the alignments from the HMM-GMM. The input layer of the DNN is composed of 15 frames (7 frames on each side of the frame for which predictions are made) where each frame corresponds to 40 log Mel-filterbank coefficients. The DNN is used to provide the posterior probability in the proposed framework for the 3450 senones defined by a decision tree.

The data used for training the UBM baseline system include only the English telephone speech of the NIST SRE, Fisher, and Switchboard sets, described in [15]. The frontend for this system extracts 20 MFCC coefficients (including C0), augmented with first order derivatives only to speed up the experiments without too much degradation. A 2048 diagonal component UBM is trained in a gender-dependent fashion, along with a 400 dimensional i-vector extractor. The dimensionality of the i-vectors is further reduced to 300 by LDA, followed by length normalization and PLDA.

The same features are used in the DNN system to compute sufficient statistics from the frame alignment given by the DNN. The i-vector, LDA, PLDA dimension and paramters are the same as for the baseline system.

As the DNN system effectively use 3450 classes, another UBM/i-vector system is trained with a 4096 diagonal component UBM for comparison. Results for a third baseline, replacing the standard UBM with a supervised UBM obtained as described in Section 5, are also shown.

Table 1 presents the performance of the baseline and proposed systems on NIST SRE'12 conditions 2 and 5. System performance is reported in terms of detection cost function (DCF) with different effective priors, equal error rate (EER), and the false alarm rate at a miss rate of 10% (FA@M10). The effective priors P^{tar} of DCF are 0.001 and 0.01 defined in NIST SRE'12 [16].

In both conditions, the supervised UBM with 3450 Gaussians performs similarly to the unsupervised UBMs. On the other hand, we observe very large improvements using our proposed DNN/i-vector approach across all measurements for both conditions. In condition 2, the proposed DNN-based approach provided roughly 25 – 35% relative improvement on $\mathrm{P}^{tar}_{0.001}$, $\mathrm{P}^{tar}_{0.01}$, and EER, and 70% on FA@M10. More surprisingly, similar improvements are observed on condition 5 as well although the DNN is trained on clean data only. The results clearly confirm the importance of the posterior estimation in the i-vector model.

Table 1. *The performance of three baseline systems compared to the proposed DNN/i-vector approach. The UBMs with 2048 and 4096 Gaussians are trained using the standard EM approach while the Gaussians in the 3450-component UBM are defined in a supervised manner by the ASR senones as described in Section 5.*

a. NIST SRE'12 C2 extended condition - female

System	$\mathrm{P}^{tar}_{0.001}$	$\mathrm{P}^{tar}_{0.01}$	EER(%)	FA@M10
UBM-EM(2048)	0.348	0.193	1.99	0.13
UBM-EM(4096)	0.333	0.184	1.81	0.11
UBM-sup(3450)	0.375	0.211	2.10	0.18
DNN	0.254	0.139	1.39	0.04

b. NIST SRE'12 C5 extended condition - female

System	$\mathrm{P}^{tar}_{0.001}$	$\mathrm{P}^{tar}_{0.01}$	EER(%)	FA@M10
UBM-EM(2048)	0.421	0.252	2.84	0.36
UBM-EM(4096)	0.401	0.237	2.55	0.26
UBM-sup(3450)	0.451	0.272	2.94	0.44
DNN	0.291	0.177	1.92	0.10

7. CONCLUSION AND FUTURE WORKS

In this work, we propose a novel scheme for speaker recognition that provides large improvements over current state-of-the-art technology by replacing the traditional UBM-GMM paradigm. The framework tightly integrates speech recognition in the speaker modeling process by using a DNN trained for phone recognition instead of a UBM-GMM to produce frame posteriors for the computation of i-vectors. The DNN produces posteriors for tied triphone state classes determined by a standard ASR decision tree. Sufficient statistics for i-vector computation are then extracted using these posteriors followed by a state-of-the-art backend which remains unchanged. This allows the system to factor out content information by comparing speakers over the same phonetic units, an approach similar to that taken in forensic comparisons.

We show the DNN approach significantly improved the i-vector speaker recognition system as compared to the traditional UBM-GMM approach in two NIST SRE'12 conditions. At a miss rate of 10%, the relative improvements in false alarm rate are on the order of 70% to 80% relative to the state-of-the-art systems, and the equal error rate decreases by 30% relatively.

The success of the proposed approach opens up a wide range of research directions for speaker recognition. The innovation of the machine learning and speech recognition community in deep learning can be easily ported over to the field of speaker recognition including tools such as convolutional neural networks, language model decoding, multi-style training, and so on. Moreover, an implicit speech recognition step for speaker recognition also opens up research to gain more insight in the understanding of the influence of speech content for speaker recognition.

8. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. ASLP*, vol. 19, pp. 788–798, May 2010.

[2] S.J.D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV-11th*. IEEE, 2007, pp. 1–8.

[3] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[4] G.E. Dahl, Dong Yu, Li Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. ASLP*, vol. 20, pp. 30–42, 2012.

[5] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Speaker verification using text-constrained Gaussian mixture models," in *ICASSP-2002*. IEEE, 2002, pp. 677–680.

[6] B. Baker, R. Vogt, and S. Sridharan, "Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification," in *Eurospeech-2005*, 2005, pp. 2429–2432.

[7] A. Park and T. J. Hazen, "ASR dependent techniques for speaker identification," in *ICSLP-2002*, 2002, pp. 1337–1340.

[8] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Politecnico di torino's 2006 NIST speaker recognition evaluation system," in *Eurospeech-2007*, 2007, pp. 1238–1241.

[9] T. Bocklet and E. Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection," in *ICASSP-2009*. IEEE, 2007, pp. 4525–4528.

[10] Xianyu Zhao, Yuan Dong, Jian Zhao, Liang Lu, Jiqing Liu, and Haila Wang, "Variational Bayesian joint factor analysis for speaker verification," in *ICASSP-2009*, 2009, pp. 4049–4052.

[11] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. ASLP*, vol. 16, pp. 980–988, July 2008.

[12] D. A. Reynolds, T. F. Quatieri, and Dunn. R. B., "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19 – 41, 2000.

[13] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[14] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *HLT '94 Proceedings of the workshop on Human Language Technology*, 1994, pp. 307–312.

[15] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation," in *Interspeech-2013*, 2013, pp. 1981–1985.

[16] "NIST SRE12 evaluation plan," http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v11-r0.pdf.